DR ALEXIS B LYONS (Orcid ID : 0000-0003-0799-5680)
DR SHANTHI NARLA (Orcid ID : 0000-0003-0228-3262)
DR RAHEEL ZUBAIR (Orcid ID : 0000-0003-3516-9650)
ILTEFAT HAMZAVI (Orcid ID : 0000-0002-3137-5601)

Article type : Original Article

Assessment of Inter-rater Reliability of Clinical Hidradenitis Suppurativa Outcome Measures Using Ultrasonography

Running Head: Assessment of HS Outcome Measures Using Ultrasound

A.B. Lyons,[1] S. Narla,[2] I. Kohli,[1,3] R. Zubair,[4] A.F. Nahhas,[5] T.L. Braunberger[1] M.K. Joseph,[6] C.L. Nicholson,[7] G. Jacobsen[8] and I.H. Hamzavi[1]

[1]Multicultural Center, Department of Dermatology, Henry Ford Health System, Detroit, MI
[2] Department of Dermatology, St. Lukes Hospital, Easton, PA
[3]Department of Physics & Astronomy, Wayne State University, Detroit, MI
[4]Department of Dermatology, Broward Hospital, Ft. Lauderdale, FL
[5]Department of Dermatology, Beaumont Health-Farmington Hills, Farmington Hills, MI
[6]Department of Dermatology, University of Michigan, Ann Arbor, MI
[7]Department of Dermatology, Wayne State University, Detroit, MI
[8]Department of Public Health Sciences, Henry Ford Hospital, Detroit, MI

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/CED.14889

Corresponding Author: Iltefat H. Hamzavi, MD
Email: ihamzav1@hfhs.org

Funding: General Electric provided the ultrasound machine used in this study

Conflicts of interest: ABL, SN, RZ, IK are sub-investigators for Lenicura and General Electric. IHH is the President of the HS Foundation, an investigator for Lenicura and General Electric, a consultant for Incyte, and is on Abbvie Advisory Board (unpaid). AFN, TLB, MKJ, CLN, and GJ have no relevant disclosures.

What is already known about this topic?

- Ultrasound can be utilized in patients with hidradenitis suppurativa to help evaluate for subclinical disease and more accurately classify severity of disease.

What does this study add?

- Ultrasound improved inter-rater agreement in this study and should be used in conjunction with physical examination findings to evaluate disease severity to ensure uniform staging.

ABSTRACT:

Background- Hidradenitis suppurativa (HS) staging and severity is typically based upon physical examination findings which can result in misclassification of severity based on subclinical disease activity and significant variation between healthcare providers. Ultrasound (US) is an objective tool to help evaluate subclinical disease and more accurately classify severity of disease. The objective of this study was to evaluate inter-rater reliability in HS disease severity assessment using clinical and US techniques.

Methods- Twenty subjects underwent clinical evaluation of HS using clinical outcome measures including Hurley, Sartorius, HS Physician Global Assessment (HS-PGA), and Hidradenitis Suppurativa Clinical Response (HiSCR) independently by two physicians. US was subsequently performed, and clinical assessments were repeated. Intra-class correlation coefficients (ICC) were obtained to evaluate inter-rater agreement of each outcome measure before and after US.

Results- Pre- to post-US improvement in ICC was seen with the Sartorius, HiSCR nodule and abscess count, and HiSCR draining fistula count. The scores went from having "good" rater agreement for Sartorius and HiSCR nodule and abscess count and "poor" rater agreement for HiSCR draining fistula count to "excellent" rater agreement amongst these scores.
Conclusions- US improved inter-rater agreement and should be used in conjunction with physical examination findings to evaluate disease severity to ensure uniform staging.

Introduction:

Hidradenitis suppurativa (HS) is a debilitating skin disease characterized by chronic, recurrent, painful inflammatory abscesses, nodules, and sinus tracts.[1] The prevalence of the disease has been reported to be between 0.00033-4.1%, although it is estimated that the disease is often underdiagnosed and misclassified overall.[2] Currently, the staging and severity of HS is determined clinically using a variety of methods dependent upon lesion counts and extent of area involvement including Hurley staging, Sartorius, and HS Physician Global Assessment (HS-PGA) amongst others.[1,3-6] Additionally, the Hidradenitis Suppurativa Clinical Response (HiSCR) is also as an outcome measure to evaluate HS severity and improvement following treatment using the number of abscesses and inflammatory nodules (AN count) and number of draining fistulas.[7] Although techniques of clinical evaluation have been helpful in the evaluation and management of HS, these scoring systems can often underestimate disease severity due to subclinical disease, and can result in significant variation between different healthcare providers. This is particularly important because the presence of more severe disease often changes management from medical to surgical.

The use of ultrasonography is emerging as a valuable objective tool to assess the stage of HS more effectively. The Sonographic Scoring of Hidradenitis Suppurativa (SOS-HS) technique was recently developed and utilizes ultrasonographic evaluation of HS affected areas to evaluate HS severity.[8] Previous studies on the use of ultrasound in the assessment of patients with HS have demonstrated that the technique can more accurately identify HS lesions and often results in up-staging of classification of the severity of disease but have not assessed the question of whether the technique improves reliability of clinical assessments between different assessors. Consequently, this study aimed to evaluate inter-rater reliability in HS severity assessment using both clinical and ultrasound (US) techniques and sought to assess the utility of ultrasonography

in supplementing the current 'gold-standard' practices of clinical assessment alone to determine clinical staging outcome measures of HS.

Materials & Methods:

This study was approved by the Institutional Review Board at Henry Ford Hospital (IRB #11505). International Conference of Harmonization (ICH) and Declaration of Helsinki Guidelines, and Good Clinical Practice (GCP) were followed in the conduct of this study. Informed consent was obtained prior to any study procedures. Our hospital system has a large HS specialty dermatology clinic with a diverse patient population from which patients were asked to participate in the study. Subjects were included if they were 18 years of age or older, were able to understand the requirements and risks of the study, were able to provide informed consent, and had a diagnosis of HS. Subjects were excluded if they were pregnant, breastfeeding, or allergic to any components of US gel. Subjects completed one study visit in which 2 physicians separately assessed for HS severity using Hurley staging, Sartorius Score, HS-PGA, and HiSCR before and after performing high frequency US imaging (variable-frequency probes with upper frequencies of 22 MHz, LOGIQ *e*, GE Healthcare, Milwaukee, WI) to determine SOS-HS. The pre-US assessments were performed separately with only 1 physician in the room at each time. The physicians were dermatology clinical research fellows who had been extensively trained in ultrasound and the HS clinical severity outcome measurement tools. Both physicians were present for the US, but only one performed it. Both physicians had separate interpretations of the ultrasound imaging that was performed. All clinical assessments and SOS-HS were graded separately, and the scores were kept from the other grading physician throughout the duration of the study.

Statistical Analysis

Statistical analyses were performed by the Division of Biostatistics and Research Epidemiology at Henry Ford Health System. Intra-class correlation coefficients (ICC) to assess inter-rater reliability were obtained pre- and post-US for each scoring system (Hurley stage, Sartorius, HS-PGA, HiSCR, and HS-SOS). The 95% confidence interval around each ICC was calculated. Pre- to post-US improvement in the ICC was considered statistically significant if the 95% confidence intervals around their post correlation coefficients did not encompass their pre correlation coefficients. Given the sample size of 20 patients, it was determined that an ICC of

0.50 should have a 95% confidence interval of no more than +/- 0.34 while an ICC of 0.90 should have a 95% confidence interval of no more than +/- 0.09. These levels of precision were determined to be adequate for the study.

Rater agreement indicates how similarly two raters scored the patients. An ICC of 1 indicates that the raters scored the patients identically while an ICC of 0 indicates there was no similarity in how they scored the patients. Correlation coefficients below 0.40 represent "poor" agreement, between 0.40 and 0.69 represent "good" agreement, and above 0.69 represent "excellent" agreement. Pre- to post-US improvement in the ICC was considered statistically significant if the 95% confidence intervals around their post correlation coefficients did not encompass their pre correlation coefficients ($P < 0.05$).

Results:

Twenty subjects (13 women, 7 men) completed the study. However, only patients containing a complete set of scores for each outcome measure from both raters were used (N=19). One patient did not have a complete set of outcome measures due to incomplete completion of the form by 1 of the physicians inadvertently. The ICC results between raters pre- and post-US assessment are summarized in **Table 1**.

The results presented in **Table 1** indicate there was "excellent" rater agreement for Hurley stage and "poor" rater agreement for the HiSCR draining fistula count before US. The Sartorius and HiSCR AN count pre-US had "good" rater agreement. After US was performed, there was no statistical change in ICC for Hurley Stage. However, the Sartorius, HiSCR AN count, and HiSCR draining fistula count achieved "excellent' rater agreement post US with statistical significance achieved (i.e. the 95% confidence intervals around their post correlation coefficients don't encompass their pre correlation coefficients). The HS-PGA demonstrated "good" rater agreement both pre- and post-US with no significant change. In addition, the HS-SOS demonstrated "good" rater agreement.

Discussion:

HS is a chronic and debilitating disease with quality of life impairment and limited effective treatment options.[9,10] As such, it is an important and rapidly expanding area of ongoing dermatologic research, but consistent clinical outcome measures remain an obstacle. These clinical outcome measures are further limited by how they are assessed currently, solely through

physical examination. Physical examination techniques, such as visualization and palpation of HS lesions, have low sensitivity, and healthcare providers may miss deep or torturous fistulae or abscesses with one study finding clinically unrecognized fluid collections in 76% of those undergoing US evaluation.[8] It is also difficult to differentiate between a draining abscess and a draining fistula. Further, a clinically palpable lesion could correspond to either a nodule, abscess, fistula, or scar, making it difficult to fully evaluate HS severity based on physical examination alone.[8]

In this study, prior to US assessment, there was "poor" rater agreement for the draining fistula count, and the Hurley stage was the only outcome measure with "excellent" rater agreement pre-US. Lack of inter-rater reliability is a major concern in many HS clinical outcome measures such as Hurley staging, Sartorius, HS-PGA, and HiSCR. Further, a recent study by Thorlacius et al (2019) investigated inter-rater reliability between 12 international HS experts (> 10 years of experience) for several HS outcome measures and found good inter-rater reliability for Hurley staging but found very wide limits of agreement for most other outcome measures.[11] Consequently, they did not recommend any of the other outcome measures (apart from Hurley stage and Physician Global Visual Analogue Scale) for measuring clinical severity of HS. Despite the good inter-rater reliability for Hurley staging for disease severity, it often does not adequately capture disease activity. Hurley staging does not incorporate patient reported outcome components or quality of life measurements and is therefore insufficient for evaluating response to treatment (eg. one Hurley stage 3 patient might be in immense discomfort due to pain, inflammation, and drainage while another Hurley stage 3 patient might not). Thus, other outcome measures are needed to adequately assess and quantify disease activity.

Ultrasound interpretation skills may vary among raters, but despite this, raters had "good" interrater agreement for HS-SOS. In addition, US examination significantly reduced interrater variability between raters for multiple outcome measures demonstrating its utility in HS assessment. The current study demonstrated that the use of US resulted in statistically significant pre- to post-US scoring improvement in the inter-rater agreement for Sartorius, HiSCR AN count, and HiSCR draining fistula count. This was due to better visualization of HS lesions with increased ability to distinguish sinus tracts, nodules, abscesses, inflammation, and scarring. This emphasizes the utility of using US along with the physical exam to obtain more accurate clinical staging and improve inter-rater reliability. High frequency and ultra-high frequency US can

provide additional detailed information beyond what the clinical examination can provide including the presence of subclinical disease activity, response to treatment, inflammation, and depth and margins of affected areas.[12-16] Further, a multicentric study conducted in HS patients comparing Hurley staging when graded clinically and with US demonstrated intra- and inter-rater US agreements of 94.9% and 81.7%, respectively.[17] In contrast to our study, this study utilized only Hurley staging and had US performed by dermatologists with 10% of the cases reviewed by a radiologist as an external consultant.

In HS, US can detect subclinical anatomical information that may significantly alter the severity, staging, and treatment options, sometimes even from a medical to surgical approach. Loo et al. found that 56.9% of patients with HS had subclinical disease seen on US.[18] A study performed by Napolitano et al. found that 28.7% of patients were found to have more severe HS measured by US SOS-HS when compared to the clinical Hurley staging system.[19] Similarly, a study by Martorell et al. revealed that for patients diagnosed with Hurley stage I disease, staging changed to a more severe stage after evaluation with US in 44.7% of patients.[17] Another study by Lacarrubba et al found that 27% of patients had worse HS measured by US when compared to clinical assessment of HS-PGA.[20] In addition, Wortsman et al. found that US findings modified the disease management in 82% of adult patients with HS, and management was changed from medical to surgical in 24% of patients.[8] A subsequent study in children (<15 years of age) with HS revealed that US findings resulted in modification of medical management of the disease in 92% of cases.[21]

Wortsman et al. recently called for US to become a standard of care for all HS patients.[22] It was further suggested that the ideal situation to perform US would be while making a baseline examination in all HS patients and then intermittently to monitor the degree of severity. US is generally widely available in many clinical and emergency departments worldwide and is part of radiology residency training programs. Moreover, there are a growing number of publications on the use of US in HS and a number of training courses offered through international US societies such as the American Institute of US in Medicine or the European Federation of Societies for US in Medicine and Biology.[22] Training to use US for examination of patients with HS is obtainable, further supporting its utility along with clinical examination to improve inter-rater reliability.

Limitations of this study include relatively small sample size, single center study, and multiple raters as research fellowships overlapped during the study. Of note, the same set of

raters performed pre- and post-US evaluation for a given subject. In addition, US is highly user-dependent, and it may have been useful for both physicians to have performed, as well as interpreted, the ultrasound individually. The nomenclature for HS lesions visualized during ultrasound is still being developed. Further work needs to be done to provide better interpretation of sonographic features of HS. Limitations of US for evaluation of HS include the inability to detect lesions less than 0.1 mm in size, decreased resolution when using lower frequencies to image deeper lesions which can limit clear visualization of edges of deep sinus tracts in those who are obese. The strengths of this study include statistical improvement in ICCs pre- and post-US demonstrated for multiple HS scoring systems despite having several different raters.

Conclusion:

As demonstrated in this study, US can help improve inter-rater reliability for assessing HS disease activity and severity. The use of clinical grading alone often underestimates the true extent of disease. US should accompany clinical examination to decrease variation in staging and severity between providers, to provide the appropriate treatment recommendations, and to evaluate treatment response. Future studies should examine differences in treatment responses in patients who have been evaluated with US versus those who were not to determine if US evaluation affects patient outcomes. These studies should also increase the number of raters and patients to ensure the changes noted in IR reliability can be verified.

References

1. Gill L, Williams M, Hamzavi I. Update on hidradenitis suppurativa: connecting the tracts. *F1000prime reports.* 2014;6:112.
2. Miller IM, McAndrew RJ, Hamzavi I. Prevalence, Risk Factors, and Comorbidities of Hidradenitis Suppurativa. *Dermatologic clinics.* 2016;34(1):7-16.
3. Hurley H. Axillary hyperhidrosis, apocrine bromhidrosis, hidradenitis suppurativa, and familial benign pemphigus: surgical approach. *Dermatologic Surgery.* 1989;133:1506-1511.
4. Alikhan A, Lynch PJ, Eisen DB. Hidradenitis suppurativa: a comprehensive review. *Journal of the American Academy of Dermatology.* 2009;60(4):539-561; quiz 562-533.

5. Sartorius K, Lapins J, Emtestam L, Jemec GB. Suggestions for uniform outcome variables when reporting treatment effects in hidradenitis suppurativa. *The British journal of dermatology.* 2003;149(1):211-213.
6. Zouboulis CC, Del Marmol V, Mrowietz U, Prens EP, Tzellos T, Jemec GB. Hidradenitis Suppurativa/Acne Inversa: Criteria for Diagnosis, Severity Assessment, Classification and Disease Evaluation. *Dermatology.* 2015;231(2):184-190.
7. Kimball AB, Sobell JM, Zouboulis CC, et al. HiSCR (Hidradenitis Suppurativa Clinical Response): a novel clinical endpoint to evaluate therapeutic outcomes in patients with hidradenitis suppurativa from the placebo-controlled portion of a phase 2 adalimumab study. *Journal of the European Academy of Dermatology and Venereology : JEADV.* 2016;30(6):989-994.
8. Wortsman X, Moreno C, Soto R, Arellano J, Pezo C, Wortsman J. Ultrasound in-depth characterization and staging of hidradenitis suppurativa. *Dermatologic surgery : official publication for American Society for Dermatologic Surgery [et al].* 2013;39(12):1835-1842.
9. Alikhan A, Sayed C, Alavi A, et al. North American clinical management guidelines for hidradenitis suppurativa: A publication from the United States and Canadian Hidradenitis Suppurativa Foundations: Part I: Diagnosis, evaluation, and the use of complementary and procedural management. *Journal of the American Academy of Dermatology.* 2019;81(1):76-90.
10. Alikhan A, Sayed C, Alavi A, et al. North American clinical management guidelines for hidradenitis suppurativa: A publication from the United States and Canadian Hidradenitis Suppurativa Foundations: Part II: Topical, intralesional, and systemic medical management. *Journal of the American Academy of Dermatology.* 2019;81(1):91-101.
11. Thorlacius L, Garg A, Riis PT, et al. Inter-rater agreement and reliability of outcome measurement instruments and staging systems used in hidradenitis suppurativa. *The British journal of dermatology.* 2019;181(3):483-491.
12. Oranges T, Vitali S, Benincasa B, et al. Advanced evaluation of hidradenitis suppurativa with ultra-high frequency ultrasound: A promising tool for the diagnosis and monitoring of disease progression. *Skin research and technology : official journal of International*

Society for Bioengineering and the Skin (ISBS) [and] International Society for Digital Imaging of Skin (ISDIS) [and] International Society for Skin Imaging (ISSI). 2019.

13. Wortsman X, Calderon P, Castro A. Seventy-MHz Ultrasound Detection of Early Signs Linked to the Severity, Patterns of Keratin Fragmentation, and Mechanisms of Generation of Collections and Tunnels in Hidradenitis Suppurativa. *Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine.* 2019.
14. Kelekis NL, Efstathopoulos E, Balanika A, et al. Ultrasound aids in diagnosis and severity assessment of hidradenitis suppurativa. *The British journal of dermatology.* 2010;162(6):1400-1402.
15. Wortsman X, Jemec G. A 3D ultrasound study of sinus tract formation in hidradenitis suppurativa. *Dermatology online journal.* 2013;19(6):18564.
16. Wortsman X, Castro A, Figueroa A. Color Doppler ultrasound assessment of morphology and types of fistulous tracts in hidradenitis suppurativa (HS). *Journal of the American Academy of Dermatology.* 2016;75(4):760-767.
17. Martorell A, Alfageme Roldán F, Vilarrasa Rull E, et al. Ultrasound as a diagnostic and management tool in hidradenitis suppurativa patients: a multicentre study. *Journal of the European Academy of Dermatology and Venereology : JEADV.* 2019;33(11):2137-2142.
18. Loo CH, Tan WC, Tang JJ, et al. The clinical, biochemical, and ultrasonographic characteristics of patients with hidradenitis suppurativa in Northern Peninsular Malaysia: a multicenter study. *Int J Dermatol.* 2018;57(12):1454-1463.
19. Napolitano M, Calzavara-Pinton PG, Zanca A, et al. Comparison of clinical and ultrasound scores in patients with hidradenitis suppurativa: results from an Italian ultrasound working group. *Journal of the European Academy of Dermatology and Venereology : JEADV.* 2019;33(2):e84-e87.
20. Lacarrubba F, Dini V, Napolitano M, et al. Ultrasonography in the pathway to an optimal standard of care of hidradenitis suppurativa: the Italian Ultrasound Working Group experience. *Journal of the European Academy of Dermatology and Venereology : JEADV.* 2019;33 Suppl 6:10-14.

21. Wortsman X, Rodriguez C, Lobos C, Eguiguren G, Molina MT. Ultrasound Diagnosis and Staging in Pediatric Hidradenitis Suppurativa. *Pediatric dermatology.* 2016;33(4):e260-264.

22. Wortsman X. Color Doppler Ultrasound: A Standard of Care in Hidradenitis Suppurativa. *Journal of the European Academy of Dermatology and Venereology : JEADV.* 2020.

Tables

Table 1. Intra-Class Correlation Results for Rater Agreement

HS Clinical Outcome Measure	Intra-Class Correlation	95% Confidence Interval
Pre-Ultrasound Assessment		
Hurley	0.705*	0.391 to 0.872
Sartorius	0.587	0.206 to 0.813
HS-PGA	0.533	0.131 to 0.785
HiSCR AN	0.690	0.366 to 0.864
HiSCR draining fistula count	0.197	0.000 to 0.580
Ultrasound Assessment		
HS-SOS	0.626	0.265 to 0.833
Post-Ultrasound Assessment		
Hurley	0.609	0.239 to 0.824
Sartorius **	0.893*	0.751 to 0.956
HS-PGA	0.585	0.203 to 0.812
HiSCR AN count **	0.924*	0.818 to 0.969
HiSCR draining fistula count **	0.748*	0.466 to 0.892

* Indicates 'excellent rater agreement'

** Indicates pre to post-US ICC improvement that is statistically significant- significant change is indicated when the pre-ultrasound correlation coefficient doesn't encompass the 95% confidence interval around the post-ultrasound correlation coefficient

Abbreviations: hidradenitis suppurativa (HS), HS Physician Global Assessment (HS-PGA), Hidradenitis Suppurativa Clinical Response (HiSCR), abscesses and inflammatory nodules (AN count), Sonographic Scoring of Hidradenitis Suppurativa (SOS-HS)